(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 October 2002 (10.10.2002)

PCT

(10) International Publication Number
WO 02/080036 A1

(51) International Patent Classification[7]: G06F 17/30, 17/21, 17/27, 17/28

(21) International Application Number: PCT/SE02/00585

(22) International Filing Date: 25 March 2002 (25.03.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0101127-9 30 March 2001 (30.03.2001) SE

(71) Applicant *(for all designated States except US)*: **HAPAX INFORMATION SYSTEMS AB** [SE/SE]; Blekingegatan 48, S-116 62 Stockholm (SE).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **EJERHED, Eva, Ingegerd** [SE/SE]; Tomtebogatan 14, S-113 39 Stockholm (SE).

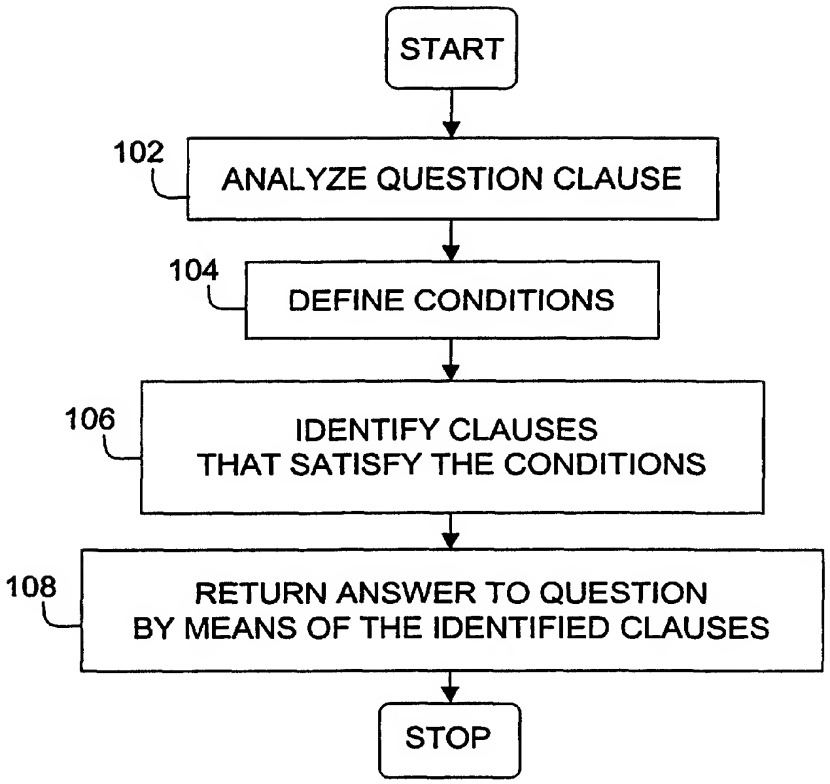
(74) Agent: **AWAPATENT AB**; Box 45086, S-104 30 Stockholm (SE).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AT (utility model), AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, CZ (utility model), DE, DE (utility model), DK, DK (utility model), DM, DZ, EC, EE, EE (utility model), ES, FI, FI (utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK (utility model), SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: METHOD OF FINDING ANSWERS TO QUESTIONS



(57) Abstract: A method and a system for automatically finding one or more answers to a natural language question in a computer stored natural language text database is disclosed. The natural language text database has been analyzed with respect to syntactic functions of constituents, lexical meaning of word tokens and clause boundaries, and the natural language question comprises a question clause. A computer readable representation of the question clause is analyzed with respect to syntactic functions of its constituents and the lexical meaning of its word tokens. In response to the analysis a set of conditions for a clause in the natural language text database to constitute an answer to the question clause is defined. The conditions relate to the syntactic functions of constituents and the lexical meaning of word tokens in the clause. Furthermore, clauses that satisfy said conditions are identified in the natural language text database, and answers to the question clause is returned by means of the identified clauses that matches the conditions.

WO 02/080036 A1

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW, ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*

— *of inventorship (Rule 4.17(iv)) for US only*

Published:

— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

METHOD OF FINDING ANSWERS TO QUESTIONS

Field of the Invention

The present invention relates to the field of information retrieval from unrestricted text in different languages. More specifically, the present invention relates to a method, and a corresponding system, for automatically finding answers to a natural language question in a natural language text database.

Background of the Invention

The field of automatic retrieval of information from a natural language text database has in the past been focused on the retrieval of documents matching one or more key words given in a user query. As an example, most conventional search engines on the Internet use Boolean search to match key words given by the user. Such key words are standardly considered to be indicative of topics and the task of standard information retrieval system has been seen as matching a user topic with document topics. Due to the immense size of the text database to be searched in information retrieval systems today, such as the entire text database available on the Internet, this type of search for information has become a very blunt tool for information retrieval. A search most likely results in an unwieldy number of documents. Thus, it takes a lot of effort from the user to find the most relevant documents among the documents retrieved, and then to find the desired information in the relevant documents. Furthermore, due to the ambiguity of words and the way they are used in a text, many of the documents retrieved are irrelevant. This makes it even more difficult for the user to find the information needed.

The performance of an information retrieval system is usually measured in terms of its recall and its precision. In information retrieval, the technical term

recall has a standard definition as the ratio of the number of relevant documents retrieved for a given query over the total number of relevant documents for that query. Thus, recall measures the exhaustiveness of the search results. Furthermore, in information retrieval, the technical term precision has a standard definition as the ratio of the number of relevant documents retrieved for a given query over the total number of documents retrieved. Thus, precision measures the quality of the search results. Due to the many documents retrieved when using the above type of search methods, it has been realized within the art that there is a need to reduce the number of retrieved documents to the most relevant ones. In other words, as the number of documents in the text database increases, recall becomes less important and precision becomes more important. Therefore, suppliers of systems for information retrieval have enhanced Boolean search by using among other things relevance ranking based on statistical methods. However, it is well known that thus highly ranked documents still comprise irrelevant documents.

Questions are a specific type of query. In the field of computerized question answering, systems range from delivering answers to simple questions to presenting complex results compiled from different sources of information. How well a question is answered is typically judged by human standards. Differently expressed, how would a well informed human being respond to a question with respect to correctness and exhaustiveness of the answer (if there is more than one answer), with respect to the succinctness of the answer to the question posed, and with respect to delivering answers quickly.

A basic difficulty for question answering systems is that, as opposed to general information retrieval systems, the inquired fact is often very specific. Thus, the need for precision becomes even more acute.

Many prior art question answering systems suffer from being dependent on knowledge specific to a domain, to a line of business or a special trade. World knowledge optimal for one domain is of little value to another and thus hard to port. To update world knowledge for a domain specific question answering system automatically is not technically feasible and such systems do not scale well.

Other prior art question answering systems that are independent of genre or domain are often restricted with regard to the type of question a user can ask, for example closed-class questions. They are direct questions whose answers are all assumed to lie in a set of objects, and are expressible as noun phrases.

Some prior art systems, such as the one disclosed in US-A-5,895,466, use language analysis in order to enhance standard keyword search. In these systems a natural language question is analyzed in order to identify appropriate keywords from the question. However, a problem with such systems is that they have low precision.

Furthermore, other prior art systems, such as the one disclosed in WO 98/25217, use complex methods of language analysis that include several representation levels and that require separate mechanisms for linking the levels to each other. The complexity of the language analysis of these systems make them unsuitable for the purpose of analyzing very large text databases.

Summary of the Invention

An object of the present invention is to provide an improved method, and a corresponding system, for automatically finding answers to a natural language question by means of a computer stored natural language text database, that are not subject to the foregoing disadvantages of existing methods for this task, i.e. that are not domain specific, that deliver answers to questions with high precision and that do not require

complex methods of language analysis. This object is achieved by a method and a system according to the accompanying claims.

The present invention is based on the insight that the relationship between the constituents and their respective syntactic functions in a question clause within a natural language question and the constituents and their respective syntactic functions in a clause that constitutes an answer to the natural language question can be used successfully in order to find answers to a natural language question in a natural language text database.

The term constituents refers to the basic units of text, such as word tokens, phrases etc. An important property of these units is that they can be found using finite state methods that recognize a strict hierarchy of constituents. Using finite state methods for syntactic analysis is well known within the art. However, the finite state method referred to here is a method of finding so-called initial clauses. Such a method is described in further detailed in the Swedish patent application SE 0002034-7 "METHOD FOR SEGMENTATION OF TEXT" and US patent application US 09/584 135 "METHOD FOR SEGMENTATION OF TEXT". Initial clauses have the property of being non-recursive, i.e. no initial clause includes another initial clause. Whenever the term clause is used in the following, it should be interpreted as initial clause.

Thus, according to a first aspect of the invention, a method is provided for automatically finding an answer to a natural language question in a computer stored natural language text database. The natural language text database has been analyzed with respect to syntactic functions of constituents, lexical meaning of word tokens, and clause boundaries, i.e. these are known to the system performing the method. The natural language question comprises a question clause, which is the clause

that conveys the content of the information need. The method comprises an analysis step, where a computer readable representation of said question clause is analyzed with respect to the syntactic functions of its constituents and the lexical meaning of its word tokens. In response to the analysis step, a set of conditions for a clause in the natural language text database to constitute an answer to the question clause is defined. The conditions relate to the syntactic functions of constituents and the lexical meaning of word tokens in the clause. Clauses that satisfy the conditions are identified in the natural language text database, and one or more answers to the question clause are returned by means of the identified clauses that satisfy said conditions.

The conditions that are defined according to the invention are based on the relationship between the constituents and their respective syntactic functions in a question clause and the constituents and their respective syntactic functions in a clause that answers the question clause. More specifically, one or more of the constituents in the question clause, or constituents that are equivalent in terms of lexical meaning, occur in a clause that answers the question, and the syntactic functions in the clause that answers the question of each of the constituents, or constituents that are equivalent in terms of lexical meaning, can be determined from the syntactic functions of the constituents of the question clause. By defining the conditions based on such relationships and then identifying clauses in the natural language text database that satisfy the conditions, an answer to a natural language question can be found without the need to rely on domain specific world knowledge. Thus, an advantage of a method of the invention is that it can be performed without the need of a large database with world knowledge which will decrease

the amount of data to store. Moreover, the precision of such a method is high.

Furthermore, the use of relations for several different type of constituents, rather than limiting the answers to a closed type and the like, also permits several answers to one question, and answers that do not necessarily identify objects by name but that still convey significant information to a user. In other words the invention identifies a limitation in prior art, where question answering systems have been considered to relate only to the answering of questions that have unique answers. In most cases this is not the case and such prior art methods thus have a limited applicability for a large set of questions (user information needs). In particular, the proposed method enables the finding of relations between persons or objects.

The term lexical meaning should be interpreted broadly. For example, in addition to word tokens that have the same lemma and word tokens that are synonyms, it is in some cases fruitful to consider word tokens that belong to the same broad semantic class to be considered as having equivalent lexical meanings. For example names, definite descriptions and personal pronouns may be interpreted as having an equivalent lexical meaning, such as the name *Jim Jarmusch*, the definite description *the director of Down by law*, and the personal pronoun *he*.

One condition in the set of matching conditions is preferably a condition relating to a lexically headed constituent having the syntactic function of main verb in the question clause. This condition stipulates that the lexically headed constituent having the syntactic function of main verb in the question clause has to have a corresponding constituent in a matching clause, i.e. a lexically headed constituent having the syntactic function of main verb and having an equivalent lexical meaning, in order for that clause to constitute an answer to the question clause. This condition introduces the use

of a condition that relates to a verb in the questions clause, which in prior art has not been considered to convey any significant information regarding the queried information.

Another condition in the set of conditions is preferably a condition relating to a lexically headed constituent having the syntactic function of subject in the question clause. This condition stipulates that the lexically headed constituent having the syntactic function of subject in the question clause has to have a corresponding constituent in a clause, i.e. a lexically headed constituent having the syntactic function of subject and having an equivalent lexical meaning, in order for that clause to constitute an answer to the question clause.

Yet another condition in the set of conditions is preferably a condition relating to a lexically headed constituent having the syntactic function of object in the question clause. This condition stipulates that the constituent having the syntactic function of object in the question clause has to have a corresponding constituent in the clause, i.e. a constituent bearing the syntactic function of object and having an equivalent lexical meaning, in order for that clause to constitute an answer to the question clause.

Moreover, further conditions on other constituents in clauses may be added to the set of conditions in order to increase the precision further. Such conditions are for example conditions relating to constituents having the syntactic functions of manner adverb, place adverb, time adverb, and causal adverb, respectively, of the question clause, or conditions relating to constituents bearing any other syntactic function. Also these condition are preferably used in combination with one or more of the other conditions.

Other syntactic functions which could be used in stating conditions are for example head and modifier.

Using such functions it is possible to find clausal answers that are expressed as noun phrases that are nominalizations of clauses. As an example the question *What did the company use to solve the problem?* can be answered by *The company used a new method to solve the problem.* but it can also be answered by the noun phrase *the company's use of a new method to solve the problem….*

The conditions above may be used separately, but they are preferably combined in such a way that they jointly state necessary and sufficient conditions for a database clause to constitute an answer to a given question clause. This increases the precision of the method even further.

In addition to, or instead of, the conditions above relating to the syntactic functions of constituents, there can be conditions only on the co-occurrence of certain constituents in a clause. For example, a condition regarding the constituents in the question clause may be defined stipulating that the constituents of the question clause, or constituents that have equivalent lexical meanings, should occur in a clause of the natural language text database in order for that clause to constitute an answer to the question clause.

Furthermore, conditions referring to a sequence of two or more clauses in the natural language text database are also envisaged.

One embodiment of the invention is directed to constituent questions (wh-questions) comprising an interrogative pronoun, such as *what*, *who*, *which* etc. According to this embodiment, i.e. where there is an interrogative pronoun in the question clause, the syntactic function of the queried constituent of the question clause is determined not only in response to the analysis step, but also in response to the interrogative pronoun. By also taking an interrogative pronoun into consideration, conditions can be defined that increase the precision of the method even further. This is due to

the fact that the interrogative pronoun itself carries information of respective semantic classes of constituents of a clause that answers the question clause. For some interrogative pronouns the syntactic function of the queried constituent is the same syntactic function as the interrogative pronoun has. For other interrogative pronouns the syntactic function of the queried constituent will be another syntactic function than the interrogative pronoun has, but it will be possible to determine the syntactic function of the queried constituent based on the identified interrogative pronoun and the analysis in the analysis step. Furthermore, the interrogative pronoun can also be used in order to determine the broad semantic class of the queried constituent. For example, the presence of the interrogative pronoun *who* in a natural language question indicates that the queried constituent is a noun phrase denoting a person.

Another embodiment concerns yes/no questions. These questions do not comprise any interrogative pronoun. Furthermore, each constituents of a question clause in a yes/no question has a corresponding constituent, i.e. a constituent that has the same lexical meaning and the same syntactic function, in a clause that constitutes an answer to the question clause. The way that a yes/no question can be distinguished from a statement will differ depending on the language. For example in some language it can be determined from the word order.

In either of the embodiments above the answer to the question may be found in a clause that satisfies the conditions. Thus, by extracting portions of text comprising the clauses that satisfy the conditions and presenting them to a user, the answer to the question clause will be evident to the user. In the embodiment concerning yes/no questions, a yes or no answer can alternatively be derived automatically from the clauses

that satisfy the conditions, and then presented to the user.

According to a second aspect of the invention, a system is provided for automatically finding an answer to a natural language question by means of a computer stored natural language text database. The system comprises storage means for storing the natural language text database. The natural language text database has been analyzed with respect to syntactic functions of constituents, lexical meaning of word tokens, and clause boundaries. The system also comprises analyzing means for analyzing a computer readable representation of a question clause of a natural language question with respect to syntactic functions of its constituents and lexical meaning of its word tokens, and defining means for defining, in response to an analysis performed by the analyzing means, a set of conditions for a clause in the natural language text database to constitute an answer to the question clause. The conditions relate to syntactic functions of constituents and lexical meaning of word tokens in the clause. The defining means are operatively connected to the analyzing means. Furthermore, the system comprises answer finding means for identifying in the natural language text database clause that satisfy the conditions and for returning an answer to the question clause by means of the clauses that satisfy the conditions. The answer finding means are operatively connected to the defining means and to the storage means.

By defining the conditions based on relationships and then identifying clauses in the natural language text database that satisfy conditions, an answer to a natural language question can be found without the need to rely on domain specific world knowledge. Thus, an advantage of the system of the invention is that the amount of data that needs to be stored is decreased and that it is possible to use the system within any domain. Moreover, the precision of the system is high.

Brief Description of the drawings

In the following, the present invention is illustrated by way of example and not limitation with reference to the accompanying drawings, in which:

figure 1 is a flowchart of a method according to an embodiment of the invention;

figure 2 is an illustration of an example of an analyzed natural language question;

figure 3A-B are illustrations of portions of text that constitute answers to the natural language question of figure 2;

figure 4 is an illustration of another example of an analyzed natural language question;

figure 5A-D are illustrations of portions of text that constitute answers to the natural language question of figure 4; and

figure 6 is a schematic diagram of a system according to an embodiment of the invention.

Detailed Description of the Invention

In figure 1 a flow chart of an embodiment of the invention is shown. In the method one or more answers to a natural language question are found in a natural language text database. One example of a natural language text database is a subset of the text information found in web servers connected to the Internet. The natural language text database has been analyzed in an antecedent process thereby enabling the use of linguistic properties of the text database in order to find answers to a natural language question. The analysis comprises the determination of a morpho-syntactic description for each word token of the natural language text database, a classification of the broad semantic class for each word token, the location of phrases in the natural language text database, the determination of a phrase type for each of the phrases, and the location of clauses in the

natural language text database. The morpho-syntactic description comprises a part-of-speech and an inflectional form, and the phrase types comprise different types according to the syntactic functions of the phrases and the part of speech of their heads. The syntactic functions comprise subject, object, main verb, adverbs etc. A clause can be defined as a unit of information that roughly corresponds to a simple proposition, or fact.

Furthermore, the natural language text database has also been indexed and stored. The spaces between each word token are numbered consecutively, whereby the location of each word token is uniquely defined by the numbers of the two spaces it is located between in the natural language text database. The interval defined by these two numbers form a unique word token location identifier. Alternative schemes for locating word tokens are known by persons skilled in the art, and the choice of which scheme to use is not critical to the invention. Since each word token is associated with a word type, it is sufficient to store all of the word types of the natural language text database and then, for each of the stored word types, store the word token location identifier of each word token associated with this word type. Furthermore, the location of a phrase is uniquely defined by the number of the space preceding the first word token of the phrase and the number of the space succeeding the last word token of the phrase. These two numbers form a phrase location identifier. Thus, each phrase type is stored and the phrase location identifier of each of the phrases of this phrase type is stored. Note that, due to the way the phrase location identifier is defined, it is easy to find out whether a word token occurs in a phrase of a certain type by determining whether the word token location identifier is included in a phrase of this type. The location of a clause is uniquely defined by the number of the space preceding the

first word token and the number of the space succeeding the last word token of the clause. These two numbers form a clause location identifier. Each of the clause location identifiers is stored. Location identifiers for sentences, paragraphs, and documents are formed in an equivalent manner and each of them is stored.

A natural language question that is to be answered in this embodiment has been classified in a prior process either as a constituent question or a yes/no question. Furthermore, the question clause of the natural language question has been identified in a prior process as well. The question clause is the clause of the natural language question that conveys the content of the information need. In a direct question, the question clause is the main clause, and in an indirect question the question clause is a subordinate clause.

In step 102 a question clause is analyzed in the same way that the natural language text database has been analyzed, i.e. the syntactic function of its constituents and the lexical meaning of its word tokens are determined. Based on this analysis, a set of conditions for a clause in the natural language text database to constitute an answer to the question clause are defined in step 104. The conditions are that at least one of the constituents in the question clause should have corresponding constituents in the clause, i.e. constituents that each has the same syntactic function and an equivalent lexical meaning as the corresponding constituent in the question clause.

When the conditions have been defined, clauses that satisfy the conditions are identified in the natural language text database in step 106 of figure 1. In the identification, the word type of the natural language text database that correspond to a word token in the question clause, and that have a lexical meaning equivalent to the word tokens in the question clause, are identified. Then the word token location identifiers

associated with the identified word types are identified in the index. The identified word token location identifiers are then used to identify the word tokens in the natural language text database that are included in a phrase of the same type as the word token in the question clause is included in, i.e. a phrase that has the same syntactic function. This is done by searching the phrase location identifiers associated with the phrase type that the word token in the question clause is included in, and determining which of the identified word token location identifiers are included in one of these phrase location identifiers. This comparison is done for each of a subset of the word tokens in the question clause, and in addition to determining if the word token is included in the same phrase type, it is determined whether the word tokens are included in the same clause. This can be done easily by determining whether the word token location identifiers are included in the same clause location identifier.

When all the clauses that satisfy the set of conditions have been identified in step 106, portions of text that each comprises one of the clauses that satisfy the set of conditions are extracted in step 108 of figure 1. These portions of text may then be presented to a user as an answer to the natural language question, or be further processed.

In the following two examples of analyzed natural language questions will be given with reference to figure 2-5. In the examples a number of abbreviations will be used which are explained in the table below:

Abbreviation	**Description**
AT	Article
NNS	Plural noun
NP	Proper noun
VB	Verb, base form
VBG	Verb present participle, gerund

VBD	Verb, past tense
WPS	Wh-pronoun, subject
WPO	Wh-pronoun, object
nps	Subject noun phrase
npo	Object noun phrase
vp	Verb phrase
cl	Clause
s	Sentence

Figure 2 illustrates an example of an analyzed natural language question. The question is: *Who is expelling diplomats?*. The question only includes one clause and the clause also constitutes a sentence. The question clause of the question is the entire question. The question clause has been analyzed with respect to a morpho-syntactic description for each word token, a lexical description (not shown) comprising lemma, a broad semantic class for each word token and synoyms, the location of phrases, a phrase type for each of the phrases, and the location of clauses. Thus, for each word token, the morpho-syntactic code is indicated, and for each space between the word tokens the number of the space is indicated. Furthermore, the location of phrases and their respective type is also indicated. Based on this analysis a set of conditions is defined for a clause in an analyzed natural language text database to constitute an answer. The natural language text database has been analyzed with respect to a morpho-syntactic description for each word token, lemma and a broad semantic class and a synonym set for each word token, the location of phrases, a phrase type for each of the phrases, the location of clauses, and the location of sentences. In this case *who* is the subject noun phrase, *expelling* is the main verb, and *diplomats* is the object noun phrase of the question clause. This will give the conditions that there should be a subject noun phrase in the clause, the lemma of the main verb in the clause

should be *expel*, and the lemma of the head of the object noun phrase of the clause should be *diplomat*, respectively, in order for the clause to constitute an answer to the question. In addition to the condition that there should be a subject noun phrase, the result of the analysis of the question clause indicates that the subject noun phrase is the queried constituent. Furthermore, the interrogative pronoun *who* indicates that this subject noun phrase should denote a person. Note that the conditions may be relaxed so that they are satisfied not only for word tokens with the same lemma, but also for word tokens that are synonyms. For example the lemma of the main verb would be allowed to be *deport* in addition to *expel*.

Turning now to figure 3A-B, portions of text that constitute answers to the natural language question of figure 2 are illustrated. The answers have been extracted from the analyzed natural language text database. In figure 3A a sentence is illustrated that includes an answer clause. In this case the first clause of the sentence has the main verb *expelling*, the object noun phrase *Russian diplomats* and the subject noun phrase *the US*. Thus, the clause satisfies the conditions above. In this case the entire sentence that the clause is included in is extracted and presented as an answer. In figure 3B a sentence is illustrated including only one clause. The clause has the main verb *expelling*, the object noun phrase *a matching number of US diplomats* and the subject noun phrase *Russia*. Thus, the clause satisfies the conditions above, and the clause is extracted and presented as an answer.

Figure 4 illustrates an example of an analyzed natural language question. The question is: *What did the ECB do?*. As in the question depicted in figure 2 the question clause of the question is the entire question. The question clause has been analyzed with respect to a morpho-syntactic description for each word token, lemma

and a broad semantic class for each word token (not shown), the location of phrases, a phrase type for each of the phrases, and the location of clauses. Thus for each word token the morpho-syntactic code is indicated, and for each space between the word tokens the number of the space is indicated. Furthermore, the location of phrases and their respective type is also indicated. Based on this analysis a set of conditions for a clause in an analyzed natural language text database to constitute an answer is defined.

The natural language text database has been analyzed with respect to a morpho-syntactic description for each word token, a broad semantic class for each word token, the location of phrases, a phrase type for each of the phrases, the location of clauses, and the location of sentences. In this case *the ECB* is the subject noun phrase, and *do* is the main verb of the question clause. The fact that *the ECB* is the subject noun phrase will give the condition that the head of the subject noun phrase in a clause should be *the ECB* in order for the clause to constitute an answer to the question. In addition to this, the interrogative pronoun *what* together with the main verb *do*, i.e. do_what, indicates that the queried constituent is an active verb phrase. Thus, a further condition is that a clause should include an active verb phrase in order for the clause to constitute an answer to the question.

Turning now to figure 5A-D, portions of text that constitute answers to the natural language question of figure 4 are illustrated. In figure 5A, a sentence including clause boundaries within the sentence illustrates one answer to the question in figure 4. In this case the first clause of the sentence has the subject noun phrase *the ECB* and an active verb phrase *has made mistakes*. Thus, the clause satisfies the conditions described with reference to figure 4. In this case the entire sentence that the clause is included in is

extracted and presented as an answer. In figure 5B, a sentence including clause boundaries within the sentence illustrates a second answer to the same question. In this case the second clause of the sentence has the subject noun phrase *the ECB* and an active verb phrase *imposed*. Thus, the clause satisfies the conditions described with reference to figure 4. In figure 5C, a sentence including clause boundaries within the sentence illustrates a third answer to the same question. In this case the first clause of the sentence has the subject noun phrase *the ECB* and an active verb phrase *has never pursued a pure policy of minimising the rate of inflation*. Thus, the clause satisfies the conditions above. Furthermore, the second clause also comprises an active verb phrase *has taken a much more practical approach of maximising the rate of growth*, but it does not include a subject noun phrase including *the ECB* and thus it does not satisfy the conditions. However, in this case the entire sentence that the first clause is included in has been extracted and presented as an answer. Thus, the relation between the active verb phrase in the second clause and *the ECB* in the first clause will be apparent to a user. In figure 5D, a sentence including only one clause illustrates a fourth answer to the same question. The clause has the subject noun phrase *the ECB* and an active verb phrase *has performed almost spectacular well*. Thus, the clause satisfies the conditions above, and the clause is extracted and presented as an answer.

Turning now to figure 6, a schematic diagram of a system according to an embodiment of the invention is shown. The system comprises analyzing means 602 for analyzing a computer readable representation of a clause, storage means 604 for storing an analyzed natural language text database, a question manager 606, defining means 610 for defining conditions for a clause to constitute an answer to a question clause, answer finding means 612 for finding clauses in a text database that

constitutes answers to a question clause, and result managing means 620. The text analyzing unit 602 is arranged to analyze a natural language text input, such as a natural language question or a natural language text database. The analysis includes the determination of a morpho-syntactic description for each word token of the natural language input, a classification of the broad semantic class for each word token, the location of phrases in the natural language input, the determination of a phrase type for each of the phrases, and the location of clauses in the natural language input. The morpho-syntactic description comprises a part-of-speech and an inflectional form, the lexical description of a word type comprises lemma, semantic class, and synonyms, and the phrase types comprises different types denoting the syntactic functions of the phrases, such as subject noun phrase, object noun phrase, other noun phrases and prepositional phrases.

In figure 6, the memory means 604, operatively connected to the text analysis unit 602, are arranged to store a natural language text database that has been analyzed by the text analysis unit 602. The natural language text database is stored in an index in the storage means 604. The indexing is based on a numbering scheme where the spaces between each word token are numbered consecutively. An alternative numbering scheme where each word token is consecutively number is also within the scope of the invention. Each word token is then defined by its word type and the numbers of the two spaces it is located between in the natural language text database. The two numbers of the spaces between which a word token is located form a word token location identifier for this word token. Furthermore, a phrase is uniquely defined by its phrase type and the number of the space preceding the first word token of the phrase and the number of the space succeeding the last word token of the phrase. The number of the space preceding the first

word token of a phrase and the number of the space succeeding the last word token of the phrase form a phrase location identifier for this phrase. Similarly, a clause, a sentence, a paragraph and a document location identifier, respectively, is defined as the number of the space preceding the its first word token and the number of the space succeeding its last word token. The word types, word token location identifiers, phrase types, phrase location identifiers, clause location identifiers, paragraph location identifiers, sentence location identifiers and document location identifiers are stored in the index that is operatively connected to the indexer. The logical and hierarchical structure of the index is shown in the table below:

Text Unit	**Location Identifiers** <i,j>
word type 1	Word token location identifiers
word type 2	Word token location identifiers
…	
word type n	Word token location identifiers
nps	Subject noun phrase location identifiers
npo	Object noun phrase location identifiers
npx	Predicate noun phrase location identifiers
pp	Preposition phrase location identifiers
cl	Clause location identifiers
s	Sentence location identifiers
p	Paragraph location identifiers
doc	Document location identifiers

Furthermore, the question manager 606 in figure 6 is operatively connected to the text analysis unit 602 and comprises defining means 610 for defining conditions for

a clause in the natural language text database to constitute an answer to a question clause that has been analyzed in the text analysis unit 602. The conditions are that a subset of the constituents in the question clause, should have corresponding constituents in the clause, i.e. constituents that each has the same syntactic function and an equivalent lexical meaning as the corresponding constituent in the question clause. Furthermore, the question manager 806 comprises answer finding means 812 for finding clauses in the natural language text database that constitutes answers to the question clause. The answer finding means 612 use the structure of the index in order to do identify clauses that satisfy the condition defined by the defining means 610. By determining the word type of a word token in a question clause, the corresponding word type in the index, and other word types in the index that have an equivalent lexical meaning give the word token location identifiers since these are stored in the index. Furthermore, since the phrase type that the word token of the question clause is included in, and the phrase type that the word tokens of the natural language text database are included in has been determined in the text analysis unit, it can be determined which of the identified word token location identifiers are included in a phrase of the same type as the word token in the surface variant, i.e. that has the same syntactic function. This is done by searching the phrase location identifiers associated with the phrase type that the word token in the question clause is included in, and by determining which of the identified word token location identifiers are included in one of these phrase location identifiers. This comparison is done for a subset of the word tokens in the question, and in addition to determining whether the word token is included in the same phrase type, the index is also used to determine whether the word tokens are included in the same clause.

Finally, in figure 6, the system comprises a result manager 612, operatively connected to the storage means 604, for extracting each portion of text comprising a clause that satisfied the conditions that are defined by the defining means. The portion of text to be extracted can be chosen as the clause satisfying the conditions, the sentence that the clause is included in, or the paragraph that the clause is included in, or the document that the clause is included in. The extraction means use the index to find the desired units (clause, sentence, paragraph or document) by consulting the respective location identifiers in the index.

CLAIMS

1. A method of automatically finding one or more answers to a natural language question in a computer stored natural language text database, wherein said natural language text database has been analyzed with respect to syntactic functions of constituents, lexical meaning of word tokens, and clause boundaries, and wherein said natural language question comprises a question clause, comprising the steps of:

analyzing a computer readable representation of said question clause with respect to syntactic functions of its constituents and the lexical meaning of its word tokens;

defining, in response to the analysis step, a set of conditions for a clause in said natural language text database to constitute an answer to said question clause, said conditions relating to the syntactic functions of constituents and the lexical meaning of word tokens in said clause;

identifying clauses in said natural language text database that satisfy said conditions; and

returning answers to said question clause by means of the identified clauses that matches said conditions.

2. The method according to claim 1, wherein said set of conditions in the defining step comprises:

a verb condition stipulating that a clause constitutes an answer to said question clause if a lexically headed constituent having the syntactic function of main verb of said question clause has a corresponding lexically headed constituent in said clause bearing the syntactic function of main verb and having an equivalent lexical meaning.

3. The method according to claim 1 or 2, wherein said set of conditions in the defining step comprises:

a subject condition stipulating that a clause constitutes an answer to said question clause if a lexically headed constituent having the syntactic function of subject of said question clause has a corresponding lexically headed constituent in said clause having the syntactic function of subject and having an equivalent lexical meaning.

4. The method according to claim 1, 2 or 3, wherein said set of conditions in the defining step comprises:

an object condition stipulating that a clause constitutes an answer to said question clause if a lexically headed constituent having the syntactic function of object of said question clause has a corresponding lexically headed constituent in said clause having the syntactic function of object and having an equivalent lexical meaning.

5. The method according to any one of the preceding claims, wherein said set of conditions in the defining step comprises:

a manner adverb condition stipulating that a clause constitutes an answer to said question clause if a lexically headed constituent having the syntactic function of manner adverb of said question clause has a corresponding lexically headed constituent in said clause having the syntactic function of manner adverb and having an equivalent lexical meaning.

6. The method according to any one of the preceding claims, wherein said set of conditions in the defining step comprises:

a place adverb condition stipulating that a clause constitutes an answer to said question clause if a lexically headed constituent having the syntactic function of place adverb of said question clause has a corresponding lexically headed constituent in said clause

having the syntactic function of place adverb and having an equivalent lexical meaning.

7. The method according to any one of the preceding claims, wherein said set of conditions in the defining step comprises:

a time adverb condition stipulating that a clause constitutes an answer to said question clause if a lexically headed constituent having the syntactic function of time adverb of said question clause has a corresponding lexically headed constituent in said clause having the syntactic function of time adverb and having an equivalent lexical meaning.

8. The method according to any one of the preceding claims, wherein said set of conditions in the defining step comprises:

a causal adverb condition stipulating that a clause constitutes an answer to said question clause if a lexically headed constituent having the syntactic function of causal adverb of said question clause has a corresponding lexically headed constituent in said clause having the syntactic function of causal adverb and having an equivalent lexical meaning.

9. The method according to any one of the preceding claims, wherein there is an interrogative pronoun in said question clause, further comprising the step of:

determining the syntactic function of the queried constituent of said question clause in response to the analysis step and said interrogative pronoun.

10. The method according to claim 9, wherein the syntactic function of the queried constituent of said question clause is determined as the syntactic function of said interrogative pronoun.

11. The method according to claim 9 or 10, wherein the analysis of lexical meaning of word tokens comprises an analysis of the broad semantic class of each word token of said natural language text database, and wherein the broad semantic class of the queried constituent is determined in response to the interrogative pronoun.

12. The method according to any one of the preceding claim, further comprising the step of:

extracting from said natural language text database portions of text comprising clauses satisfying said conditions.

13. A system for automatically finding one or more answers to a natural language question in a computer stored natural language text database, comprising:

storage means comprising said natural language text database which has been analyzed with respect to syntactic functions of constituents, lexical meaning of word tokens, and clause boundaries;

analyzing means for analyzing a computer readable representation of question clause of a natural language question with respect to syntactic functions of its constituents and lexical meaning of its word tokens;

defining means, operatively connected to said analyzing means, for defining, in response to an analysis performed by the analyzing means, a set of conditions for a clause in said natural language text database to constitute an answer to said question clause, said conditions relating to the syntactic functions of constituents and the lexical meaning of word tokens in said clause; and

answer finding means, operatively connected to said storage means and said defining means, for identifying in said natural language text database clauses that satisfy said conditions and for returning answers to said

question clause by means of said clauses that satisfy said conditions.

14. A computer readable medium having computer-executable instructions for a general-purpose computer to perform the steps recited in any of the claims 1-12.

15. A computer program comprising computer-executable instructions for performing the steps recited in any of the claims 1-12.

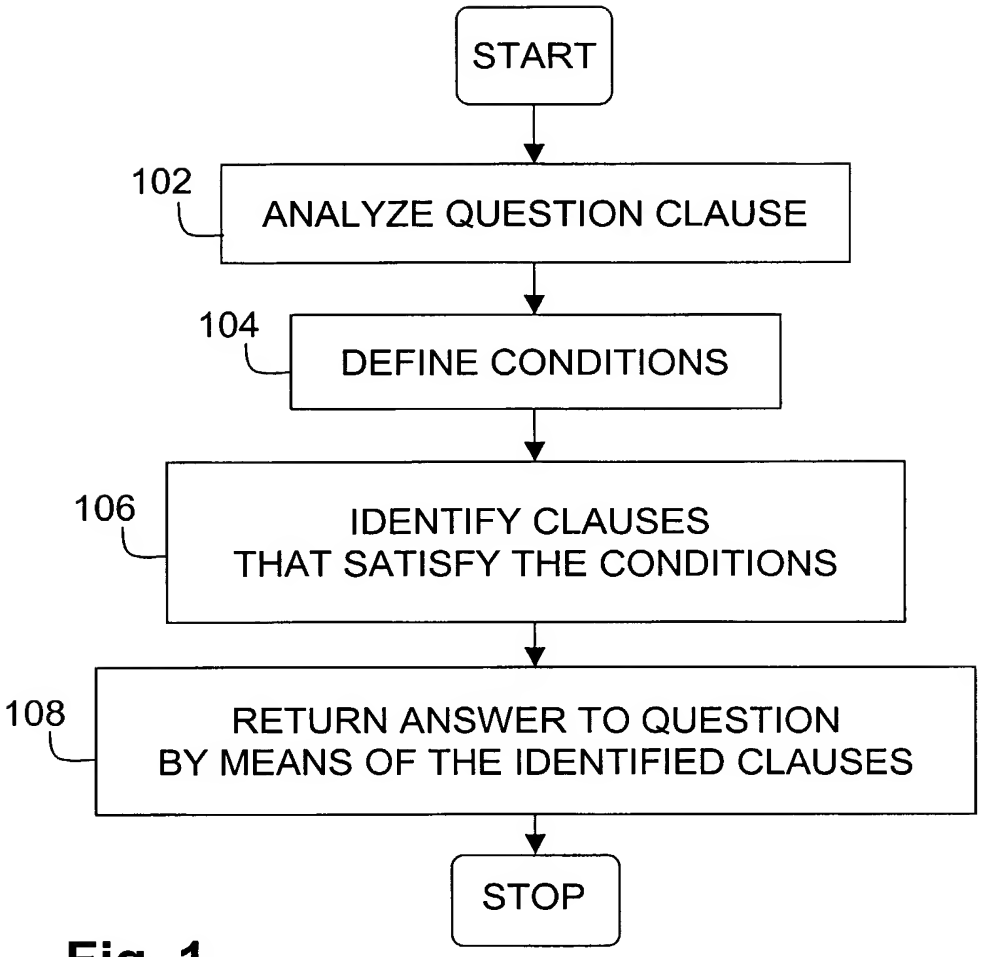

START
102
ANALYZE QUESTION CLAUSE
104
DEFINE CONDITIONS
106
IDENTIFY CLAUSES
THAT SATISFY THE CONDITIONS
108
RETURN ANSWER TO QUESTION
BY MEANS OF THE IDENTIFIED CLAUSES
STOP

Fig. 1

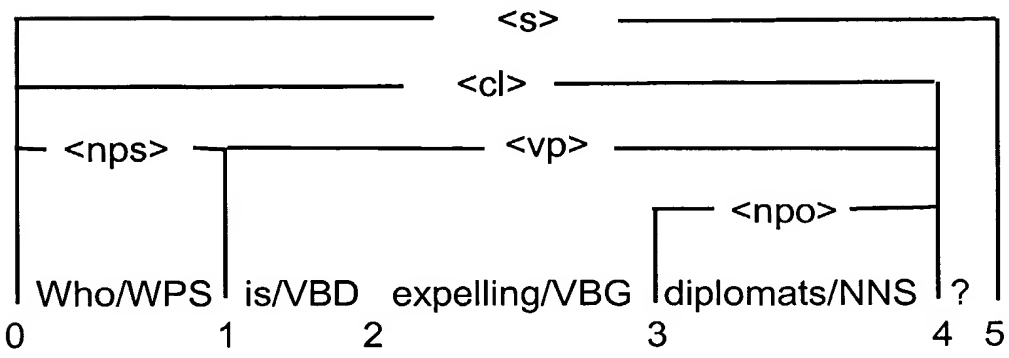

<s>
<cl>
<nps>
<vp>
<npo>
Who/WPS
is/VBD
expelling/VBG
diplomats/NNS
?
0
1
2
3
4
5

Fig. 2

<cl1>*The US* is expelling 50 Russian diplomats, including four said to have "run" Robert Hanssen, the FBI agent accused last month of spying for Russia</cl1>.

Fig. 3A

<cl1> *Russia* is expelling a matching number of US diplomats </cl1>.

Fig. 3B

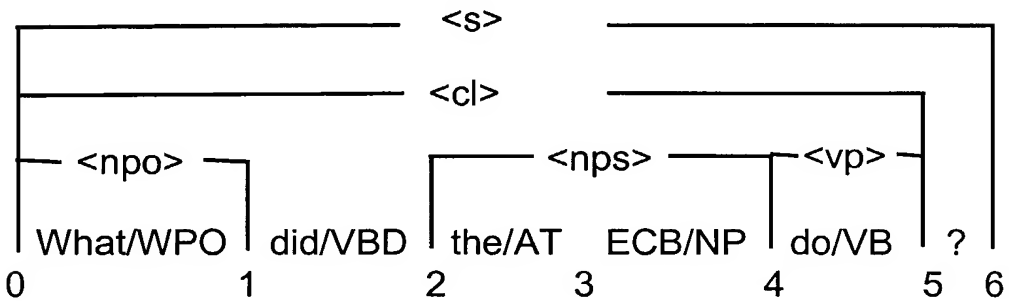


Fig. 4

<cl1> The ECB *has made mistakes*, </cl1> <cl2> and has been frequently criticized for them</cl2>.

Fig. 5A

<cl1> The euro-zone inflation rate will soon fall below the tolerance level of 2 per cent </cl1> <cl2> that the ECB *imposed* </cl2>.

Fig. 5B

<cl1> The ECB *has never pursued a pure policy of minimising the rate of inflation*, </cl1> <cl2> but has taken *a* much more practical approach of maximising the rate of growth, given an acceptable rate of price increases </cl2>.

Fig. 5C

<cl1>The ECB *has performed almost spectacularly well*, much better certainly than the Bundesbank in its dying years, and better even than the Federal Reserve during the recent boom and bust </cl1>.

Fig. 5D

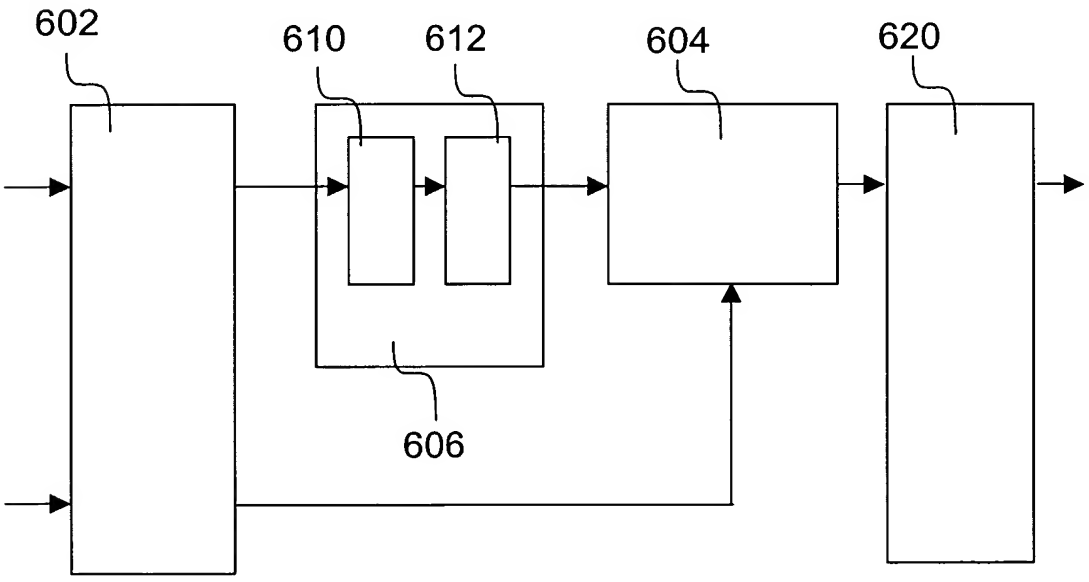


Fig. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 02/00585

A. CLASSIFICATION OF SUBJECT MATTER

IPC7: G06F 17/30, G06F 17/21, G06F 17/27, G06F 17/28
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC7: G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
-	US 5895466 A (GOLDBERG, R.G. ET AL), 20 April 1999 (20.04.99) --	1-15
-	WO 9825217 A (QUARTERDECK CORPORATION), 11 June 1998 (11.06.98) --	1-15

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 June 2002	05 -07- 2002
Name and mailing address of the ISA/ Swedish Patent Office Box 5055, S-102 42 STOCKHOLM Facsimile No. +46 8 666 02 86	Authorized officer Oskar Pihlgren /OGU Telephone No. +46 8 782 25 00

Form PCT/ISA/210 (second sheet) (July 1998)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 02/00585

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.: 1-15
because they relate to subject matter not required to be searched by this Authority, namely:
see extra sheet

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

Information on patent family members

10/06/02

International application No.
PCT/SE 02/00585

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5895466 A	20/04/99	CA 2244826 A	19/02/99
WO 9825217 A	11/06/98	AU 5381698 A	29/06/98
		EP 1016003 A	05/07/00